Finding missing references in learning courses

Patrick Siehndel, Ricardo Kawase, Asmelash Teka and Eelco Herder
L3S Research Center
Leibniz Universität Hannover, L3S, Appelstr. 9a, 30167 Hannover, Germany
{siehndel, kawase, teka, herder}@L3s.de

ABSTRACT

Reference sites play an increasingly important role in learning processes. Teachers use these sites in order to identify topics that should be covered by a course or a lecture. Learners visit online encyclopedias and dictionaries to find alternative explanations of concepts, to learn more about a topic, or to better understand the context of a concept. Ideally, a course or lecture should cover all key concepts of the topic that it covers, but often time constraints prevent complete coverage. In this paper, we propose an approach to identify missing references and key concepts in a corpus of educational lectures. For this purpose, we link concepts in educational material to the organizational and linking structure of Wikipedia. Identifying missing resources enables learners to improve their understanding of a topic, and allows teachers to investigate whether their learning material covers all necessary concepts.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*User Issues*

General Terms

Experimentation, Verification

Keywords

Linked Data, Wikipedia, Education

1. INTRODUCTION

The availability of linked data for a growing set of topics offers us new opportunities in the way we can use this structured information. In this paper we explore and analyze methods for finding missing content in educational material by exploiting the links and categories used in Wikipedia.

For teachers and authors of learning material it is a challenge to select and cover the key concepts that belong to this topic, while keeping the time required to study the material within certain limits. This selection process is guided by the teacher's - partially subjective - point of view on the topic, the intended learning goals and the prerequisite knowledge that learners are assumed to have. As a result, learning material may suffer from missing references that are either central to the given topic or that are required to understand certain parts of the learning material.

Many websites and projects aim to leverage learning through technology enhancing learning tools, such as online platforms that provide courses, lectures, tools and community communication among others. Coursera[1], Udacity[2], OpenCourseWare[3] are a few examples of such efforts. One organization in particular caught much attention from media as well as from TEL community: the Khan Academy[4] educational organization. In 2009 it receive the Microsoft Tech Award for education, followed by a $2 million support from Google for the creation of more courses and translation of content in 2010[5]. Much of Khan Academy's success is attributed to its low-tech, but high quality conversational tutorials, with lessons that are quick, free, and easy to understand.

Apart from dedicated courses, many learning activities are carried out using search engines, dedicated blogs and reference websites, among which Wikipedia[6] is the most popular[7]. Additionally, the debates around learning topics and exercises through the means of discussion forums, social networks or even e-mails shifted the learning process from a paper-based activity and solitary task to a Web-based[3] and collaborative activity. These 'non-educational' resources and the rich interlinking in sites such as Wikipedia provide a good coverage of topics, but do not provide learners the focus and preselection of material that is available in educational resources.

In this paper, we aim to bridge this gap between educational material and non-educational resources by identifying resources that may need to be included or referenced in on-

[1] http://www.coursera.org
[2] http://www.udacity.com
[3] http://www.ocwconsortium.org
[4] http://www.khanacademy.org
[5] http://www.google.com/campaigns/project10tothe100
[6] http://www.wikipedia.org
[7] http://www.ebizmba.com/articles/reference-websites

Table 1: Mapping of required LOM fields.

Khan Academy Topics	Wikipedia Categories	URL
Algebra	Algebra	http://en.wikipedia.org/wiki/Category:Algebra
Applied Math	Applied mathematics	http://en.wikipedia.org/wiki/Category:Applied_mathematics
Arithmetic and Pre-Algebra	Arithmetic	http://en.wikipedia.org/wiki/Category:Arithmetic
Art History	Art History	http://en.wikipedia.org/wiki/Category:Art_history
Biology	Biology	http://en.wikipedia.org/wiki/Category:Biology
Calculus	Calculus	http://en.wikipedia.org/wiki/Category:Calculus
Chemistry	Chemistry	http://en.wikipedia.org/wiki/Category:Chemistry
Geometry	Geometry	http://en.wikipedia.org/wiki/Category:Geometry
Healthcare and Medicine	Health Care	http://en.wikipedia.org/wiki/Category:Health_care
History	History	http://en.wikipedia.org/wiki/Category:History
Physics	Physics	http://en.wikipedia.org/wiki/Category:Physics

line educational material. Given the fact that Wikipedia is the biggest and most accessed reference website (almost 9 billion page views per month with over 4 million articles[8]), our goal is to identify relevant references that could improve and support the learning of given subject. As a simple example, reading the *Pythagorean theorem*[9] can be of great benefit for someone who is following Geometry lectures[10] at the Khan Academy.

As the Khan Academy is not compliant with Linked Data standards [2], our work requires a first enrichment step in which the courses are annotated with mentions to Wikipedia articles, i.e. learning references. Our work focuses on uncovering a strategy that can identify relevant missing references in lectures, after the enrichment is done (or given any enriched dataset). The benefits of uncovering missing references are twofold: first, learners are able to better understand a lecture by studying relevant references not explicitly cited in the corpus and further deepen their knowledge in a given topic; second, teachers and educators are able to discover what might be further explored or what has mistakenly been overlooked.

The rest of this paper is organized as follows. In Section 2, we investigate existing work on the field. In Section 3, we briefely introduce the contents of the Khan Academy and the data preparation for our research. In Section 4, we present the different strategies for identifying missing references in Wikipedia. Section 5, exposes our user studies where we validate the most appropriate strategies, followed by our conclusions in Section 6.

2. RELATED WORK

In this section, we position our work in the context of related literature, organized as (i) the missing link problem, (ii) linking free text to Wikipedia articles and (iii) computing semantic relatedness using Wikipedia.

The closest project to our work is the 2008/9 Wikipedia selection for schools[11]. This project launched by SOS Children UK and the Wikimedia Foundation[12] compiled manually selected Wikipedia articles for school children on various topics. The content can be navigated using a pictorial subject index, or a title word index. This has the advantage that it is clean, however it is not scalable and it is not easy to generate links to similar pages.

Automated approaches for recommending missing links to related articles in Wikipedia have been proposed in [1, 5, 10]. In [5], the authors proposed a topic-model based approach for recommending missing links to related articles by harnessing the link text of Wikipedia articles. Given an article, they compute the similarity of its topic distribution with other articles, using this relation, they provide the recommendation of related articles for the input article. In [1] the authors use clustering based on co-citation and page title information of an input Wikipedia article to rank related articles to it. Then, they collect anchor text from outgoing links of the related articles to see if any of them are missing in the input page.

Another line of research related to our work deals with linking free text to Wikipedia articles. Linking unstructured data to Wikipedia articles has been studied in [6, 7, 8] among others. In [7] the authors use Wikipedia as a resource for automatic keyword extraction and word sense disambiguation. They provide a system, Wikify!, that automatically identifies important words and phrases in text and links them to their corresponding Wikipedia articles. Similarly, [8] uses machine learning to identify significant terms within unstructured text, and enrich it with links to the appropriate Wikipedia articles. [6] provides a system for automatically annotating text documents with DBpedia[13] URIs. In our work, we use [8] to disambiguate key phrases from Khan Academy class video transcriptions to identify their corresponding Wikipedia articles.

Finally, we look into related work that tries to measure semantic relatedness using Wikipedia. In [9] the authors use Wikipedia's hierarchical category structure to measure the semantic relatedness of terms. In [11] the authors use the hyperlink structure of Wikipedia for obtaining measures of semantic relatedness. In [4] authors propose Explicit Semantic Analysis, ESA, a method that represents the meaning of texts in high dimensional vector using Wikipedia concepts. In our work, we look into the category and link structure of Wikipedia to quantify semantic relatedness and to recom-

[8] http://stats.wikimedia.org/EN/Sitemap.htm

[9] http://en.wikipedia.org/wiki/Pythagorean_theorem

[10] http://www.khanacademy.org/math/geometry

[11] http://schools-wikipedia.org/

[12] http://bit.ly/XAyPIf

[13] http://dbpedia.org/

mend related articles.

3. KHAN ACADEMY

Khan Academy is a non-profit educational organization and a website created in 2006 by Salman Khan. The goal of the Khan Academy is to provide high quality education for anyone, anywhere. Up to date, the website provides a free online collection of over 4,000 micro lectures[14]. The lessons are in video format, all of them hosted via YouTube and available within the Khan Academy website.

The lessons cover several topics, including mathematics, history, health care, medicine, finance, physics, chemistry, biology, astronomy, economics, cosmology, and organic chemistry, American civics, art history, macroeconomics, microeconomics, and computer science. In addition to the videos, the website also supports different features such as progress tracking, practice exercises, and a variety of tools for teachers in public schools. The leactures are narrated in English and most of them have an interactive transcript.

As previously mentioned, we employ the Kahn Academy's dataset to investigate concepts that might be missing in a course. In order to do that, we crawled all video leactures with available transcripts. In total, we collected 2,283 transcripts with an average length of 1,045 words.

4. APPROACH

The first step in our approach consists of discovering the existing links in Khan Academy's lectures to Wikipedia references. As previously mentioned, Khan Academy is not compliant with Linked Data standards, making any semantic analysis unfeasible. Therefore, we first annotate lectures' scripts to detect any mention of entities that can be linked to Wikipedia articles. For this purpose, we use the WikipediaMiner [8] service as an annotation tool. The WikipediaMiner approach consists of two basic steps: first, detected words are disambiguated using machine learning algorithms that take the context of the word into account.

This step is followed by the detection of links to Wikipedia articles: only those words that are relevant for the whole document are linked to articles. The goal of the whole process is to annotate a given document in the same way as a human would link a Wikipedia article. Our Wikipedia dataset contains over 4 million articles covering almost all knowledge domains. In order to identify all existing links, we set the confidence parameter to the lowest value possible. In total, the process generated 170,465 annotations to 18,275 unique Wikipedia references.

4.1 Category Mapping

The second step in our work consists of accurately contextualizing the annotations. Khan Academy employs a three-level course structure for organizing fields of study, subjects and topics. For example, in the field of study *Math* there are subjects such as *Algebra*, *Geometry* and *Calculus*. Further, within *Algebra* there are topics such as *Linear Equations*, *Functions*, and *Matrices*. We manually assessed the subjects in order to align them with Wikipedia categories. This helped us to identify contextualized references (found in the annotation process) and in addition serves as one subgraph building strategy (see Subsection 4.2). The mapping is exposed in Table 1.

4.2 Finding Relevant Articles

Based on the category mapping, we extend the context of a learning subject by expanding the references graphs. We analyzed three different ways on how to build a subgraph for a given category.

Direct Category (Simple) This is the basic strategy to build a subgraph for a given topic. In this approach, we take the articles which are directly related (mapped) to the given category. Thus, this strategy will only suggest references that are directly associated to the Wikipedia category that is aligned with Khan Academy's topic. The Wikipedia categories were manually related to each given Khan Academy topic. Table 1 shows which Kahn Academy topics are mapped to which Wikipedia categories.

SubCategory Building the graph based on the subcategories of the given main category increases the size of the resulting graph on the cost of adding irrelevant articles. Instead of taking just articles which belong directly to the given category, we also consider articles which belong to the Wikipedia's subcategories of the given category. Depending on how many levels of subcategories are parsed, one can control the size of the resulting tree. The subcategories of a given category are in most cases relatively close to the parent category in terms of covered topics. For example, the subcategories of 'Algebra' are 'Theorems in Algebra', 'Elementary Algebra', 'Linear Algebra', to name but a few. In most cases, subcategories cover a special topic of the parent category.

Outlink This strategy starts with the articles which are related to the main category and adds all articles which are mentioned as outlinks in one of these articles. Similar to the subcategory based approach, we can control the size of the resulting graph by limiting the number of outlink levels that are taken into account. Exploiting outlinks increases the size of the resulting graph much faster than the previous approach. Additionally, the topics covered by the articles in the resulting graph are much broader and less related to the original topic. For instance the Article 'Algebra' in Wikipedia links to many topic very close related to 'Algebra' but due to the fact that also 'History of Algebra' is described in the article, references such as 'Alexandria' or 'Greeks' are linked as well.

Figure 1 gives an overview of the different strategies and shows which articles are added based on the different strategies. The figure limits to depict the first level of each strategy. The second level for the subcategory based approach would, for instance, take into account all articles which belong to the subcategories of 'Universal algebra', 'Variables', 'Polynomials' and 'Elementary algebra'.

[14] http://www.khanacademy.org/about

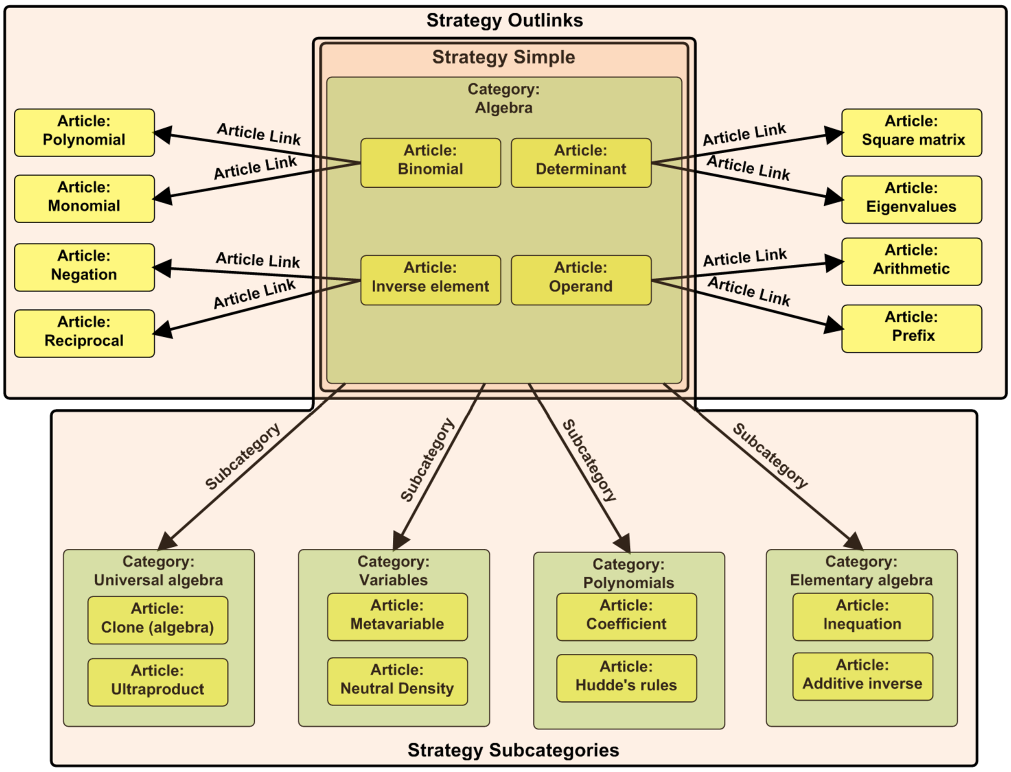


Figure 1: Graph Construction

Table 2: Graph Size

Graph Strategy	Avg. Number of Articles per Category
Simple	128.91
Outlink 1st Level	2877.00
Outlink 2nd Level	85182.73
Subcategory 1st Level	2136.00
Subcategory 2nd Level	9879.54
Subcategory 3rd Level	31990.55

The number of contextualized articles (possible suggestions of missing references) strongly diverge based on the chosen strategy. Table 2 shows the number of articles that are inside the resulting graph for each different strategies.

In the Kahn Academy, taking only the direct mapping into account, the average number of articles related to the main categories is 129. The most conservative strategy, which extends the graph based on subcategories produces a graph 16 times bigger. If we consider all articles which are reachable by taking into account outlinks for two levels, we get more than 85,000 articles per topic.

Obviously, not all articles in the resulting graphs are relevant to a given learning topic. In order to select and present only relevant articles, we tried three different strategies for ranking the set of articles. The strategies and features we used for ranking are:

Wikipedia Inlinks The articles which have the highest number of incoming links are selected to be the most relevant. One can assume that articles which are more often linked have a high relevance for many topics, therefore these article should be covered by a given topic.

Wikipedia Outlinks This strategy is based on the assumption that articles which link to many other articles are relevant because they act as a hub. Additionally, the high number of outgoing links represents a higher human effort (Wikipedia editors) in explaining the article. Thus, suggesting that this given article is more elaborated, more important and possibly, for us, more relevant.

Subgraph Inlinks For this strategy we computed the number of inlinks to the given articles only considering articles inside the newly created graph. The idea behind this strategy is that articles which are more often linked to inside the created graph (stronger connected) play a significant role for the given graph. Additionally, since the graph is created based on the topic of interest we assume that taking the subgraph inlink counting may reveal closer related articles for the given topic.

In order to get an overview how the different strategies cover the topics discussed inside Kahn Academy, we performed a preliminary analysis of the generated graphs. We selected different sets of representative articles from the Kahn

Academy courses. The representativeness of an article was calculated based on the relevance (the relevance of an article for the given text is provided by WikipediaMiner) and the number of courses in which it was found.

For calculating the precision and recall we started by taking the 100 most representative elements from Kahn Academy and the same amount of elements from each strategy ordered by the number of inlinks in the subgraph. In cases where a strategy suggested less than 100 elements, we took all elements in consideration. Based on this setup we got relatively poor results for precision and recall. The best performing strategy was the one based on outlinks (1st level), with a recall and precision of 0.25. A closer look at the results revealed that, by just taking 100 elements from each strategy we are not considering the characteristics of each algorithm. The simple strategy is supposed to deliver a few good quality results. Thus, by taking a fixed set of 100 elements we also take very low ranked results into account, caused by the low number of elements the method suggests. In contrast to this, the other strategies produce a much bigger set of elements which are not necessarily all mentioned in Kahn Academy, but might still be relevant for the topic. Additionally, we expected that the strategies which take more elements into account should cover a bigger set of elements from the Kahn Academy and therefore get a higher recall.

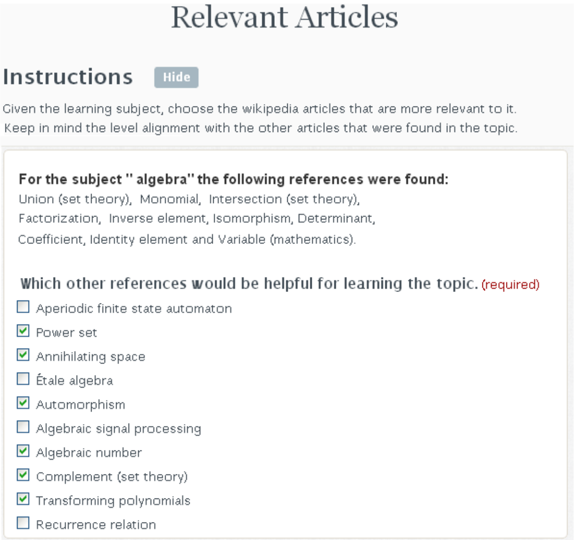
For analyzing this in detail, we decided to take the top 20 percent of the elements, again based on the subgraph inlink ordering strategy (with a maximum of 5,000 articles). By doing so we increased the number of elements for all other strategies and reduced the number of elements from the simple strategy. Based on this setup we got a very high precision of 0.77 for the simple strategy, which indicates that the relatively small number of suggested elements were very relevant for the topic. By contrast, the outlink (2nd level) based approach got a precision of 0.1 but a recall of 0.8. The best performing strategy based on the f-measure was the outlink based approach with just one level, with an f-measure of 0.32. Since the goal of the approach is to find missing elements, we performed a user study where we analyzed the usefulness of the suggested references not covered in the Kahn Academy.

5. USER STUDY

In order to evaluate the quality and utility of the suggestions, we set up a user evaluation to collect assessments of the results. The goal is to validate which combination of article selection and ranking provides best references to a learning topic.

The evaluation follows this setup: first, an evaluator is presented with the title of a topic of study (see Table 1) and the top ten Wikipedia references identified in the transcripts of the lectures. In this way, the evaluator can have an overview of what are the themes covered by a given topic.

In addition to that, the evaluator is presented with a list of ten additional Wikipedia articles that are provided by one of the strategies from Section 4. This list is composed by the top five and the bottom five articles of a given strategy. The items are randomly positioned in a multiple choice interface (check boxes) to avoid biased judgments. The evaluators must choose the items that they believe to be most relevant and aligned (in terms of complexity) to the topic. There is no minimum or maximum limit of choices. The evaluator might choose none, some or all the articles.



Relevant Articles

Instructions

Given the learning subject, choose the wikipedia articles that are more relevant to it.
Keep in mind the level alignment with the other articles that were found in the topic.

For the subject " algebra" the following references were found:
Union (set theory), Monomial, Intersection (set theory), Factorization, Inverse element, Isomorphism, Determinant, Coefficient, Identity element and Variable (mathematics).

Which other references would be helpful for learning the topic. (required)

- ☐ Aperiodic finite state automaton
- ☑ Power set
- ☑ Annihilating space
- ☐ Étale algebra
- ☑ Automorphism
- ☐ Algebraic signal processing
- ☑ Algebraic number
- ☑ Complement (set theory)
- ☑ Transforming polynomials
- ☐ Recurrence relation

Figure 2: Evaluation interface.

Implicitly, all items should be relevant to the given topic due to the nature of the subgraphs, especially for the simplest graph that is solely based on the topic-category mapping. However, for the other subgraph strategies, the relevance most likely decreases as the graph grows. Therefore, the setup of this evaluation enables us to access the most suitable strategy (how many articles are chosen) and the most suitable ranking feature (how many of the top articles are chosen). Figure 2 depicts the evaluation interface that was set up in CrowdFlower[15].

5.1 Results

In total, we used 12 combinations of subgraph strategies and ranking. Applied to each of the 11 learning topics, this results in 132 unique evaluations that we manually accessed.

We had three expert evaluators that volunteered to participate in the study. The results are summarized in Table 3. The results should be interpreted as follows: The third column (average number of items chosen) regards specifically how good a subgraph strategy is to find related articles in a learning topic (values range from 0 to a maximum of 10); the fourth column (average number of top items) indicates how well the ranking strategy performs (values can range from 0 to 5 and are limited to the average number of items chosen). Higher values in the *top items* column indicate that the ranking strategies were adequate for the subgraph strategies. Lower values indicate that the evaluators' choices came from the *bottom* of the ranking, which in principle represents a random selection.

In this sense, the *simple* graph strategy that represents the

[15] https://www.crowdflower.com

Table 3: Evaluation results. All combinations of the Simple strategy, and the top performing combinations for the remainder strategies.

Graph Strategy	Ranking	Avg. number of items chosen	Avg. number of top items
Simple	Outlinks	4.3636	2.9091
Simple	Inlinks	4.0909	3.2727
Simple	Subgraph Inlinks	**4.8182**	3.1818
SubCategories Lvl1	Inlinks	3.7273	2.8182
SubCategories Lvl2	Subgraph Inlinks	3.9091	3.7273
Outlinks Lvl1	Subgraph Inlinks	3.8182	2.3636

direct mapping of learning topics to Wikipedia categories, combined with *subgraph inlinks* ranking, is the best performing one. Evaluators chose in average 4.8 Wikipedia articles that are suitable references to a given learning topic.

Additionally, we see that in most cases, ranking based on the *number of inlinks* performs better. For example, in the combination *SubCategories Lvl1 + Outlinks*, ranking plays a minor role since the top ranking choices occur in less than 60% of the cases (2.3636). On the other hand, *inlinks* provide much better results, as in the noteworthy case of *SubCategories Lvl2 + Subgraph Inlinks*, where over 95% of the references chosen belong to the top ranking list.

6. CONCLUSIONS

In this paper, we dealt with the problem of identifying missing relevant references in educational lectures. We explored several strategies to build a relevant network of references, combined with different ranking methods. Our results show that a simple mapping of learning subjects to Wikipedia categories provides the most relevant results. In addition, exploring first level of subcategories also leads to quality suggestions with higher diversity. On the contrary, the results also suggest that the article linking structure of Wikipedia is not able to support either contextualization of topics or relevancy. In addition, *inlink* strategies for ranking were, without dispute, the best approaches to choose appropriate related references.

Our approach can be applied to any textual resource, provided that the annotation step is performed. We believe that results can be further improved if the data is manually annotated or compliant with Linked Data principles. The implications of our work are beneficial for both learners and educators. Learners are able to deepen their knowledge and improve the understanding on different subjects by studying these references, while educators can be informed about further topics that should be taught.

7. REFERENCES

[1] S. F. Adafre and M. de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 90–97, New York, NY, USA, 2005. ACM.

[2] T. Berners-Lee. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[3] M. A. Chatti and M. Jarke. The future of e-learning: a shift to knowledge networking and social software. *Int. J. Knowledge and Learning*, 3 (4/5):404–420, 2007.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, 2007.

[5] C. Haruechaiyasak and C. Damrongrat. Article recommendation based on a topic model for wikipedia selection for schools. In G. Buchanan, M. Masoodian, and S. Cunningham, editors, *Digital Libraries: Universal and Ubiquitous Access to Information*, volume 5362 of *Lecture Notes in Computer Science*, pages 339–342. Springer Berlin Heidelberg, 2008.

[6] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

[7] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[8] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.

[9] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[10] O. Sunercan and A. Birturk. Wikipedia missing link discovery: A comparative study. In *AAAI Spring Symposium on Linked Data Meets Artificial Intelligence (Linked AI 2010), ser. AAAI Spring Symposium, AS Symposium, Ed., Stanford, USA*, 2010.

[11] I. H. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.